RESEARCH PROPOSAL

An Enhanced Privacy Preserving Technique for Cloud Computing Environment

For registration to

Doctor of Philosophy

IN THE FACULTY OF SCIENCE

to

IIS (Deemed to be a university), JAIPUR

Submitted by:

Kamlesh Ahuja
Enroll. No.: ICG/2017/24798

Under the Supervision of:

Dr. Navneet Sharma
Associate Professor
IIS (Deemed to be a university)

Department of CS & IT

September 2018

An Enhanced Privacy Preserving Technique for Cloud Computing Environment

Introduction

Cloud computing is one of the popular computational infrastructures due to scalability, reliability, and ability of computation. Cloud computing can be used for computational needs as well as the data storage. In this presented work the cloud based privacy preserving data mining framework is proposed work design and investigation. The data mining is a technique to explore and understand the data and their hidden pattern which may helpful for making decisions, predictions and other applications. The data mining techniques are basically recognized as supervised and unsupervised in nature. In a supervised learning technique, the learning patterns are provided with their predefined classes. On the other hand, the unsupervised learning algorithms trained by self. Therefore, the supervised learning techniques are known as classification techniques and unsupervised learning techniques are recognized as the clustering technique. Basically, the data mining techniques are employed on a centralized database or data sources [1].

But now in these days, the requirement of data analysis and pattern detection is extended for multiple data sources. In addition to that in this environment, the security and privacy of data become an essential concern [2]. Basically, in multiple data source mining the data is collected from multiple applications and the need to combine the data for finding the application oriented outcomes but no one party which are participating in data mining want to disclose the data and the generated outcomes. Therefore, the nature of data is being changed in addition to that the application of data is also changed with respect to the traditional data mining techniques.

Such kind of data mining technique is known as the privacy preserving data mining technique [3]. In this environment, the basic requirement is to secure the sensitive part of data during the analysis of data and the discloser of the results. in addition to that sometimes such data is also utilized in different applications for research purpose, therefore, the data utilization and their distribution is also a key aim of the privacy preserving data mining. In this presented work in

place of traditional data mining techniques the domain of privacy preserving data mining and their issues are intended to explore.

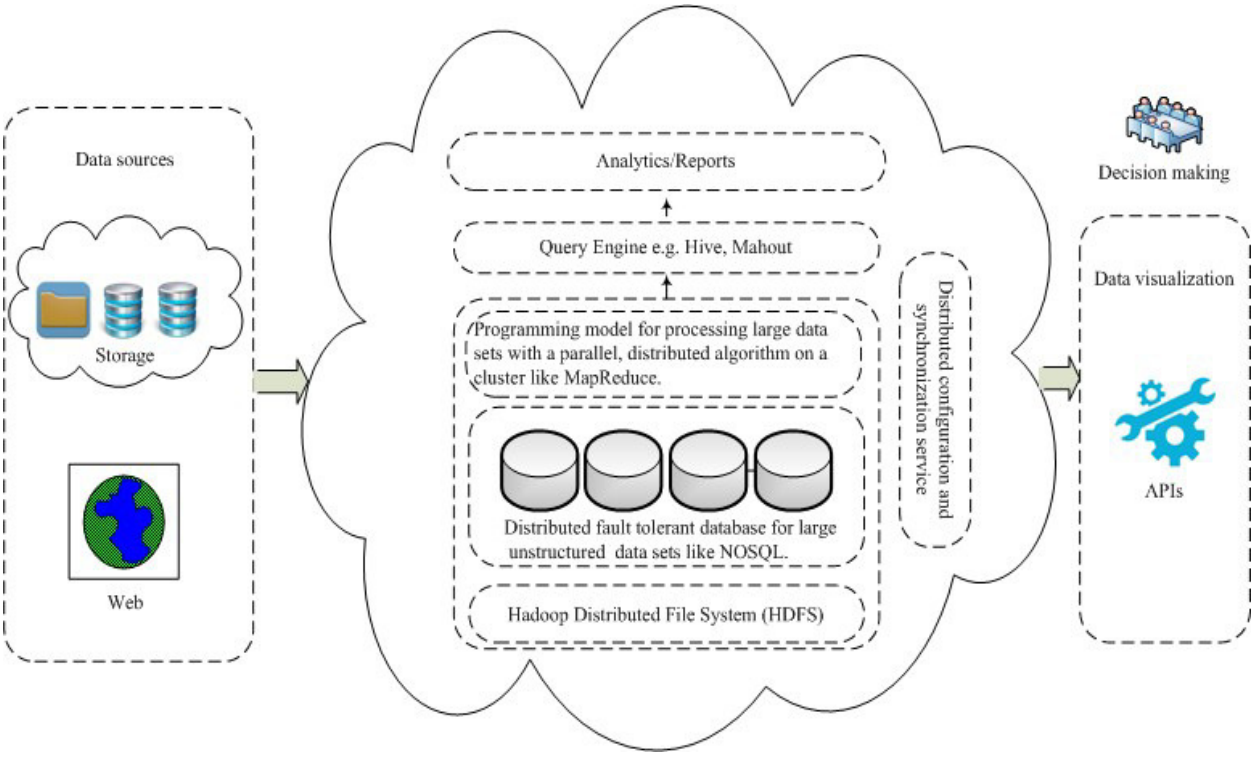
Data sources
Storage
Web
Analytics/Reports
Query Engine e.g. Hive, Mahout
Programming model for processing large data sets with a parallel, distributed algorithm on a cluster like MapReduce.
Distributed fault tolerant database for large unstructured data sets like NOSQL.
Hadoop Distributed File System (HDFS)
Distributed configuration and synchronization service
Decision making
Data visualization
APIs

Figure 1 Cloud Computing Usage in big data

Uses of Data Mining

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps [4]:

- **Data cleaning:** also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation:** in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

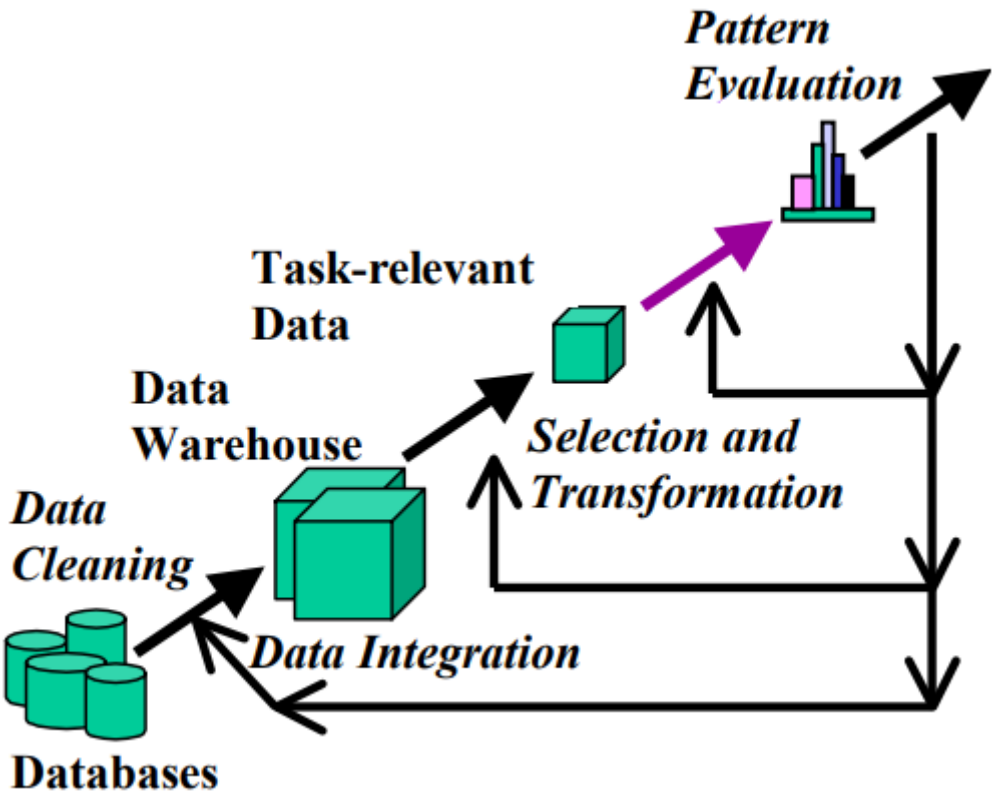

Figure 2 Knowledge Discovery in Databases process

Privacy preserving System for cloud computing

Privacy-preserving data mining considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm. The main consideration of PPDM is twofold [5]. First, sensitive raw data like identifiers, names, addresses and so on, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. So, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. The former is referred to as individual privacy preservation and the latter is referred to as collective privacy preservation [6].

- **Individual privacy preservation:** The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.
- **Collective privacy preservation:** Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups and, at the same time, should prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve strategic pattern that is paramount for strategic decisions, rather than minimizing the distortion of all statistics. In other words, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

Limitations of Privacy Preservation in Data Mining: Privacy Preservation Data Mining techniques do not mean perfect privacy, for example, The SMC computation won't reveal the

sensitive data, but the data mining result will enable all parties to estimate the value of the sensitive data. It isn't that the SMC was "broken", but that the result itself violates privacy.

Models of PPDM

In the study of privacy-preserving data mining (PPDM), there are mainly four models as follows:

1. Trust Third Party Model

The goal standard for security is the assumption that we have a trusted third party to whom we can give all data. The third party performs the computation and delivers only the results– except for the third party, it is clear that nobody learns anything not inferable from its own input and the results. The goal of secure protocols is to reach this same level of privacy preservation, without the problem of finding a third party that everyone trusts.

2. Semi-honest Model

In the semi-honest model, every party follows the rules of the protocol using its correct input, but after the protocol is free to use whatever it sees during the execution of the protocol to compromise security.

3. Malicious Model

In the malicious model, no restrictions are placed on any of the participants. Thus any party is completely free to indulge in whatever actions it pleases. In general, it is quite difficult to develop efficient protocols that are still valid under the malicious model. However, the semi-honest model does not provide sufficient protection for many applications.

4. Other Models - Incentive Compatibility

While the semi-honest and malicious models have been well researched in the cryptographic community, other models outside the purview of cryptography are possible. One example is the interesting economic notion of incentive compatibility. A protocol is an incentive compatible if it can be shown that a cheating party is either caught or else suffers an economic loss. Under the rational model of economics, this would serve to ensure that parties do not have any advantage by cheating. Of course, in an irrational model, this would not work.

We remark, in the "real world", there is no external party that can be trusted by all parties, so the Trust Third Party Model is an ideal model.

Application of privacy preserving data mining

A number of different applications and the domains exist for PPDM. Among some essential applications are reported in this section [7].

- **Credential Validation Problem:** In this problem, we are trying to match the subject of the credential to the person presenting the credential. For example, the theft of social security numbers presents a serious threat to security. In the credential validation approach, an attempt is made to exploit the semantics associated with the social security number to determine whether the person presenting the SSN credential truly owns it.
- **Identity Theft:** A related technology is to use a more active approach to avoid identity theft. The identity angel system crawls through cyberspace and determines people who are at risk from identity theft. This information can be used to notify appropriate parties. Note that both the above approaches to prevention of identity theft are relatively non-invasive and therefore do not violate privacy.
- **Web Camera Surveillance:** One possible method for surveillance is with the use of publicly available webcams, which can be used to detect unusual activity. Note that this is a much more invasive approach than the previously discussed techniques because of person specific information being captured in the webcams. The approach can be made more privacy-sensitive by extracting only facial count information from the images and using these in order to detect unusual activity. It has been hypothesized that unusual activity can be detected only in terms of facial count rather than using more specific information about particular individuals. In effect, this kind of approach uses a domain-specific downgrading of the information available in the webcams in order to make the approach privacy-sensitive.
- **Video-Surveillance:** In the context of sharing video-surveillance data, a major threat is the use of facial recognition software, which can match the facial images in videos to the facial images in a driver license database. While a straightforward solution is to completely black out each face, the result is of limited new, since all facial information has been wiped out. A more balanced approach is to use selective downgrading of the facial information so that it scientifically limits the ability of facial recognition software

to reliably identify faces while maintaining facial details in images. The algorithm is referred to as k-Same, and the key is to identify faces which are somewhat similar and then construct new faces which construct combinations of features from these similar faces. Thus, the identity of the underlying individual is anonymized to a certain extent, but the video continues to remain useful. Thus, this approach has the flavor of a k-anonymity approach, except that it creates new synthesized data for the application at hand.

- **The Watch List Problem:** The motivation behind this problem is that the government typically has a list of known terrorists or suspected entities which it wishes to track from the population. The aim is to view transactional data such as store purchases, hospital admissions, airplane manifests, hotel registrations or school attendance records in order to identify or track these entities. This is a difficult problem because the transactional data is private, and the privacy of subjects who do not appear in the watch list need to be protected. Therefore, the transactional behavior of non-suspicious subjects may not be identified or revealed. Furthermore, the problem is even more difficult if we assume that the watch list cannot be revealed to the data holders. The second assumption is a result of the fact that members on the watch list may only be suspected entities and should have some level of protection from identification as suspected terrorists to the general public. The watch list problem is currently an open problem.

Limitation of Existing PPDM approaches

Privacy preserving data mining is a vast domain of research and development. A significant number of contributions are researched are existing in privacy preserving data mining and their techniques. After reviewing existing approaches, we conclude some key research challenges:

1. **Current approaches are not supporting for re-query of data after data manipulation:** In PPDM environment more than the single party is involved in mining the data and obtain the results. But not a single one wants to disclose the sensitive records to others or mining outcome to others. Therefore, the handling all the security and privacy requirements of data and the data owner's area complicated task.

2. **Nature of data aggregator or miner:** In the PPDM scenarios a single place is required where the data is aggregated and mined for finding the patterns and trends from data. In this context it is required, where the data is aggregated and mined is secure and trustworthy. But practically there are no fully trusted data aggregator exists.
3. **Both kind of data formats vertically and horizontally is not feasible to be processed at the same time using the same technique:** Not only the security and privacy in this environment is a key concern of mining but managing the type of data and their association is also a challenge in PPDM environment. The nature of data source and handling of data sources are also a challenging task in terms of horizontally partitioned data and vertically partitioned data.
4. **Downgrading the learning performance:** In PPDM techniques for mining the data supervised or unsupervised learning techniques are utilized. In both, the techniques quality of data played an important role in finding the patterns. But for securing the sensitivity and privacy of data the noise over the data or the encryption policies are applied. That directly impact on the performance of data mining algorithms.
5. **There is not any provision to validate privacy after publishing a dataset using the PPDM approach:** When the PPDM based data mining is required a number of different data sources are participating on the data mining. The information from different data sources is aggregated on a single dataset. After that, the mining operations are conducted. In this context identification of meaningful attributes and the higher dimensions of data is a challenge for PPDM. In addition to that after mining collection and results and their distribution to concerning, parties are also a complicated task.
6. **After data publishing some of the data leakage issues are possible due to system complexity:** In PPDM environment some of the applications are available where the publishing of data sets for third party use is required in this context identification of sensitive and private part of information and their discloser is required which can be misused by the third party. Handling of such attributes in an effective manner is a remains a research direction for PPDM.

Literature review

In this section, related work is analyzed and reviewed which extends the guidelines for designing a new approach to privacy preserving data mining.

The privacy preserving data mining is newer technology as compared to the other streams (domains) of data mining. In this technology, the security of end user identity and the privacy hidden in data is the aim. In different scenarios where the data need to be a club and then generate new kind of data set for experimentation using the realistic data, the privacy preserving techniques can be utilized. In this similar scenario, ***Mahtab Arafati et al [8]*** proposes the data as a service technique to generate the realistic dataset by combining the different sources of data. not only for cloud based system the technique is also suitable for big data technology. In this context, ***Kaihe Xu et al [15]*** proposed to work in PPDM technology for big data. The resultant data can be utilized with any data mining model for finding the valuable outcomes for research and another study purpose, therefore, ***Lichun Li et al [9]*** proposes to study about the PPDM using association rule mining technique over the vertically partitioned data. In a similar manner, ***Xun Yi et al [14]*** provides the method to effectively compute the association rules in PPDM environment. In PPDM the sensitive information can disclose the identity of the end user therefore to support such kind of sensitivity on the outsourced data ***Chun-Wei Lin et al [10]*** an hiding-missing-artificial utility (HMAU) algorithm is proposed to hide sensitive item sets through transaction deletion. Not only intentions sometimes by mistake the data sensitivity remains in the data therefore before exposing the data to the third party the reevaluation of data is required. Therefore, ***Xiaokui Shu et al [11]*** present a privacy preserving data-leak detection (DLD) solution to solve the issue. During the data arrogation in a common place the data is encrypted or transformed in other formats by which the privacy issues remain fixed. But due to this data is not recovered in their actual formats thus ***Jin Li et al [12]*** propose L-EncDB, a novel lightweight encryption mechanism for databases for this issue. In the similar contribution on effective encryption algorithm development ***Qingchen Zhang et al [13]*** scheme, only the encryption operations, and the decryption operations are performed by the client. During the search queries the data only which having the exactly similar words can be recoverable in this context a new data model which works on the semantics of the input query and outcomes ***Zhangjie Fu et al [16]*** propose a content-aware search scheme. Therefore, a number of

applications of this innovative technology are become possible for example ***Juha Oksanen et al [17]*** have developed a method for generating a privacy-preserving heat map with user diversity (ppDIV) and tracking. Additionally, for IOT ***Prem Prakash Jayaraman et al [18]*** and for recommendation system **Cong Wang et al [19]** provide noteworthy applications using privacy preserving data mining techniques.

The following points are the limitations in existing approaches. Overcoming these limitations serves as a motivation in our proposed work.

The PPDM can be used for any kind of environment (i.e. cloud, big data or any central repository)

1. The different strategy required to deal with both kinds of data aggregation environment (i.e. vertical or horizontal) that depends on the nature of data and required outcomes.
2. The data dimensions are an essential factor and for finding the high performance on PPDM privacy preserving dimensionality reduction techniques or feature selection techniques required.
3. After processing data with the PPDM technique there are some chances to results some data leakage from the privacy point of view.
4. Re-query of data in some applications are required therefore need some technique by which the reverse recovery of encrypted data and the processed data can be feasible.

Objectives

This proposed work integrates the data mining techniques and cloud technology for finding the improved methodology of cloud data analysis and their privacy. As the data mining contains the three main phases pre-processing of data, learning with patterns and the publishing of data. In a similar manner, the privacy preserving techniques can also include the protection techniques in all the phases.

1. **To design a secure and privacy preserving data storage in the cloud using some cryptographic technique (AES is an efficient algorithm for cryptography and also accept different sizes of keys thus we are using the AES algorithm for security and scalability of cryptographic scenarios).**

2. **To design a privacy preserving technique that is able to process the information in cryptographic scenarios.**
3. **Implementation and deployment of proposed privacy preserving technique over the cloud environment.**
4. **Performance analysis of the proposed data model.**

Performance parameters

The PPDM techniques basically involve the data mining and the security during mining the data and their data conclusion distributions. In this context, the main aim of PPDM approaches is to mine the data and extract the data mining patterns from the data with preserving the privacy of data and owners. Here the following performance parameters are evaluated for providing the performance of the system.

1. Precision: The precision of a data mining algorithm demonstrate how accurately patterns are recognized. Using this parameter evaluation, it is concluded that is the data outcomes are different from their initial outcomes or similar as they are at initial.
2. Recall: That provides the specific subject information during the classification analysis in data mining applications. Therefore, if the patterns are changed from their initial formats how that performance parameter is changing is analyzed using this parameter.
3. F-measures: The f-measure shows the variations on the outcomes in form of precision and recall.
4. Memory utilization: The memory usages of the system is also termed as the space complexity of the algorithms thus how the resources are applied and utilized with the given model is computed using the memory utilization.
5. Time requirements: How much amount of time required for system learning and prediction is measured using time requirements. That is also known as the time complexity of the system.

Research Approach

- In phase I a survey shall be conducted to find secure and privacy preserving technique for cloud data hosting using cryptographic techniques.
- In phase II implementation of privacy preserving data mining technique for cryptographically secured data.

- In phase III implementation and extending of the technique for vertically partitioned as well as horizontally partitioned data.
- In phase IV combine the model to demonstrate the security and privacy based data mining in a cloud environment.
- In phase V the performance evaluation of the proposed techniques with the similar techniques available in the literature.

Tools and technology

The implementation required for this software and hardware on the development side system:

(A) Tools-

User Interface Design (UI Design)-Net Beans IDE 8.1

(B) Technology/Framework-

Framework-JDK 8

(C) Minimum Hardware Requirement-

3 GB storage disk

1024 MB RAM

Intel P4 or higher Processor

(D) Software Specifications-

Linux or Windows-based platform

(E) Deployment

OPENSHIFT open cloud platform

Work Plan

The work plan indicates the projected timing of the next milestone in the research. We have divided our research work into different phases which are as follows:

Phase Number	Approximate time duration (in months)	Phase Description
1	For the first 5-6 Months	Modeling of PPDM models for horizontal and vertical partitioned data
2	For the next 4-6 Months	Addressing the issue of high dimensional data and applying a dimension reduction technique on PPDM model
3	For the next 6 Months	Addressing the issue of noise over the data and implementing the noise regulation technique for optimal noise and privacy management
4	For the next 6-8 Months	Performance analysis of the proposed data model
5	At least 5-6 Months	Thesis Design

References

[1] L. A. Kurgan & Petrmusilek (2006), "*A survey of Knowledge Discovery and Data Mining process models*", The Knowledge Engineering Review, Cambridge University Press, Vol. 21:1, 1–24.

[2] A.Adhikari, P.Ramachandrarao, B. Prasad, and J.Adhikari (2010), *"Mining Multiple Large Data Sources*", The International Arab Journal of Information Technology, Vol. 7, No. 3.

[3] X. Li, Z. Yan, P. Zhang (2014), *"A Review on Privacy-Preserving Data Mining"*, IEEE International Conference on Computer and Information Technology (CIT), 10.1109/CIT.2014.135

[4] O. R. Zaïane (1999), *"Introduction to Data Mining"*, CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta.

[5] N.Kashyap, Dr. V.Bhattacharjee (2015), *"Security in Privacy Preserving Data Mining"*, International Journal Of Engineering And Computer Science, ISSN:2319-7242, Volume 4 Issue, Page No. 11698-11703

[6] R.Natarajan, Dr.R.Sugumar, M.Mahendran, K.Anbazhagan (2012), *"A survey on Privacy Preserving Data Mining"*, International Journal of Advanced Research in Computer and Communication EngineeringVol. 1, Issue 1.

[7] M.Younus, Dr.A. A.Alhamed, K. M. Farooq, F. Begum (2014), *"Data Mining Modeling Techniques and Algorithm Approaches in Privacy Data"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6.

[8] M.Arafati, G. G. Dagher, B. C. M. Fung, P. C. K. Hung (2014), *"D-Mash: A Framework for Privacy-Preserving Data-as-a-Service Mashups"*, IEEE 7th International Conference on Cloud Computing (CLOUD)

[9] L. Li, R. Lu, K. K. R. Choo, A.Datta, and J. Shao (2016), *"Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases"*, 1556-6013 (c) IEEE.

[10] C. W. Lin, T. P. Hong, and H. C. Hsu (2014), *"Reducing Side Effects of Hiding Sensitive Item sets in Privacy Preserving Data Mining"*, Hindawi Publishing Corporation, Scientific World Journal, Article ID 235837, 12 pages

[11] X. Shu, D. Yao, and E.Bertino (2015), *"Privacy-Preserving Detection of Sensitive Data Exposure"*, IEEE Transactions On Information Forensics and Security, VOL. 10, NO. 5.

[12] J. Li, Z. Liu, X. Chen, F.Xhafa, X. Tan, D. S. Wong (2014), *"L-EncDB: A lightweight framework for privacy-preserving data queries in cloud computing"*, Knowledge-Based Systems

[13] Q. Zhang, L. T. Yang, and Z. Chen (2016), "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning", IEEE Transactions on Computers, VOL. 65, NO. 5.

[14] X. Yi, F. Y. Rao, E.Bertino, A.Bouguettaya (2015), *"Privacy-Preserving Association Rule Mining in Cloud Computing"*, ASIA CCS'15, Singapore. Copyright ACM 978-1-4503-3245-3/15/04

[15] K. Xu, H. Yue, L.Guo, Y.Guo, Y. Fang (2015), *"Privacy-preserving Machine Learning Algorithms for Big Data Systems"*, IEEE 35th International Conference on Distributed Computing Systems

[16] Z. Fu, F. Huang, K. Ren, J.Weng, and C. Wang (2017), *"Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data"*, IEEE Transactions on Information Forensics and Security, VOL. 12, NO. 8.

[17] J. Oksanen, C. Bergman, J.Sainio, J.Westerholm (2015), *"Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data"*, Journal of Transport Geography 48, 135–144

[18] P. P. Jayaraman, X. Yang, A. Yavari, D. Georgakopoulos, X Yi (2017), "Privacy preserving Internet of Things: From privacy techniques to a blueprint architecture and efficient implementation", Future Generation Computer Systems, Published by Elsevier B.V.

[19] C. Wang, Y. Zheng, J. Jiang, K. Ren (2018), "Toward Privacy-Preserving Personalized Recommendation Services", Engineering 4-21–28, Published by Elsevier LTD